IMPROVING THE QUIZ
Student Preparation and Confidence as Feedback Metrics

Pantelis M. Papadopoulos[1], Antonis Natsis[1], and Nikolaus Obwegeser[2]
[1]*Centre for Teaching Development and Digital Media, Aarhus University, Aarhus, Denmark*
[2]*Department of Management, Aarhus University, Aarhus, Denmark*
{pmpapad, anatsis}@tdm.au.dk, nikolaus@mgmt.au.dk

Keywords: Feedback, Group Awareness, Formative Assessment, Quiz, Confidence, Preparation.

Abstract: The study analyzes the potential of different feedback metrics that could improve learning in quiz-based activities. For five consecutive weeks, a group of 91 sophomore students started their classes on Information Systems with a short multiple-choice quiz. The quiz activity was organized into three phases: (a) provide initial response to the questions, (b) view feedback on class activity and revise initial responses, and (c) discuss correct answers and class performance with the teacher. The feedback included information on the percentage of students that selected each choice, on students' self-reported levels of preparation, and their, also self-reported, confidence that their initial responses were correct. The students used an online quiz tool that was developed for the study and were randomly distributed into four groups, according to the type of feedback they received (only percentage; percentage & confidence; percentage & preparation; percentage, confidence, & preparation). Result analysis revealed that students were relying first and foremost on the percentage metric, even in cases where a wrong answer had the highest percentage value. However, statistical analysis also revealed a significant main effect for confidence and preparation metrics in questions where the percentage metric was ambiguous (i.e., several choices with high percentages).

1 INTRODUCTION

Quiz activities, and multiple-choice instruments in general, are widely used in different learning settings. A quiz can be used in the beginning, the middle, or the end of a class, inside and outside of the classroom, and it can be administered by the teacher or be optionally used by the student. When used in the beginning of a class, a quiz can present to the teacher a valuable picture of students' prior knowledge, making it easy to identify issues and misconception. Similarly, short quizzes during the class could act like clickers and reassure the teacher that the students are able to follow the lecture (Buil, Catalán, & Martínez, 2016), while a quiz at the end of the class could provide the opportunity for a review to the students.

In computer supported education, formative feedback could include timely, personalized, and customizable feedback (Sosa, Berger, Saw, & Mary, 2011). This, in turn, could provide additional opportunities to the student for self-reflection and self-assessment (Bransford, Brown, & Cocking, 2000; Kleitman & Costa, 2014; Wang, 2008).

There is a plethora of open and free tools available that allow the teacher to design, set up, and administer quiz activities for different learning purposes. Each tool could offer unique affordances that would better match instructional needs, but the basic premise remains selecting the correct answer out of a predefined set of choices. For example, Socrative[1] allows the teacher to monitor student progress through a series of quizzes, thus also monitoring the progress of a student throughout a semester. Quiz activities in PeerWise[2] are based on student-generated questions. The system allows the student to answer questions submitted by peers and review their quality and level of difficulty. PeerWise is also utilizing gamification, by including badges and leaderboards (Denny, 2013). Finally, Kahoot[3] allows the user to create a range of different closed-type game-like activities, such as multiple choice questions, fill-in-the-blanks, etc. The tool

[1] http://www.socrative.com/
[2] https://peerwise.cs.auckland.ac.nz/
[3] http://getkahoot.com

emphasizes its game-like characteristics, introducing also competition between the users.

This research focuses on the uses of quiz activities for formative assessment, and examines metrics that could provide a better feedback to the students, by integrating objective and subjective information in depicting class knowledge.

2 BACKGROUND

2.1 Quiz and Group Awareness

The feedback the student receives in Socrative, PeerWise, and Kahoot can be based both on information previously submitted by the teacher/designer (e.g., predetermined feedback for a wrong answer in a question) and on information related to fellow students' activity (e.g., group score, percentage of students selected each option). Regarding the latter, Bodemer (2011) suggested that comparability is an essential part of tools focusing on group awareness, arguing that allowing students to compare their knowledge with that of their peers' can be beneficial for their learning. Despite this, it is worth noting that the feedback that a student receives in these three tools stays on the surface, focusing only on the percentage of students under each alternative choice in the quiz. Although the use of the percentage metric could be easily understood and useful for the students, it cannot provide additional qualitative information that could be useful for a student, in terms of comparison and self-assessment.

Several studies have already explored the learning benefits from supporting group awareness, analyzing the desirable characteristics of group awareness tools (e.g., Janssen & Bodemer, 2013; Lin, Mai, & Lai, 2015, for a review). In general, group awareness can refer to cognitive (e.g., what do the peers know?) or social (e.g., what do the peers do in the group?) information about the group members (Buder, 2011). Since this study explores the potential of multiple-choice quizzes, the term "group awareness" refers to an aggregated view of the group knowledge, as represented through different metrics.

In the context of the study, the group refers to the whole class population, while the used metrics include, apart from the percentage metric, subjective information (i.e., peers' self-reported levels of confidence and preparation). Studies combining objective and subjective metrics have already suggested that this combination can be beneficial for the students (e.g., Erkens, Schlottbom, & Bodemer, 2016; Schnaubert & Bodemer, 2015). For example, Kleitman and Cost (2014) reported that asking students how confident they were that their answers were correct improved their metacognition. We argue that, similarly, the goal of increasing group awareness in quiz-based activities could be better served when a more detailed view of the class knowledge is offered to the students, by including both objective and subjective metrics in the feedback.

2.2 Student Learning and Engagement

Research findings have repeatedly underlined the beneficial impact quiz activities could have on students' motivation and performance. Méndez-Coca and Slisko (2013) used Socrative to engage students in active learning. Students' responses in follow-up surveys showed a wide appreciation of the approach, mentioning among others benefits that the use of Socrative made them more involved in the classes and stimulated their interaction with their peers. The latter can be easily linked to the multifold benefits of externalizing one's knowledge. Even though answering multiple-choice questions does not provide the space that a writing task would on justification, structure, and argumentation, making students' opinions explicit can provide a useful foundation for meaningful peer interaction (Papadopoulos, Demetriadis, & Weinberger, 2013). In their study (Méndez-Coca, & Slisko, 2013), teachers grouped together students with different opinions, arguing that such a pairing could promote dialogue amongst students.

Also in favor of quiz activities, students appreciate, in general, this type of learning activity. For example, DiBattista, Mitterer, and Gosse (2004) analyzed student attitudes towards multiple-choice testing with immediate feedback assessment. Even though their study focused on comparing an immediate feedback system against multiple-choice tests conducted by pen-and-paper, students' opinions were overwhelmingly positive towards the former. What is more important is that this preference for immediate feedback was not correlated to students' actual performance or their personal characteristics.

Apart from the immediate feedback a computer-supported quiz can offer, another reason of their appeal is arguably their game-like nature. In-class quiz activities often integrate gamification in the learning process (Deterding Dixon, Khaled, & Nacke, 2011). Getting the correct answer translates into points, credits, badges, better positions in a

leaderboard and so on. Although these game elements introduce rewards that are usually detached from the learning process, their impact on student engagement has been observed in several studies (e.g., Denny, 2013; Wang, 2013). It needs to be emphasized, though, that student engagement that is based on quiz's novelty effect or superficial awards may decrease over time (Wang, 2013). Gamification needs to be part of a purposeful instructional design, to avoid having students "gaming the system" (Baker, Walonoski, Heffernan, Roll, Corbett, & Koedinger, 2008) or disengaging because of the competition gamification can inject in the learning process (Papadopoulos, Lagkas, & Demetriadis, 2016).

2.3 Study Motivation

The current study discusses the impact of two metrics, in addition to the percentage one, that could better depict the knowledge level of students in the class, namely their level of preparation and their level of confidence. Both metrics are self-reported, thus subjective. The preparation metric shows how prepared the students feel, just *before* they take the quiz, while confidence is a metric indicating how sure the student is *after* having answered a question in the quiz.

Finally, it is worth noting that the current study is part of a larger research project focusing on the potential of closed-type formative assessment tools that could be easily used by the teacher to increase student engagement and performance. The research project also examines how multiple, short, quiz-based activities can provide enough information to build student knowledge profiles and how these profiles can be later used and affect direct collaboration activities that occur in the course (e.g., group project assignments). Nevertheless, the discussion on the long-term outcomes of this project spans outside the scope of this paper.

3 METHOD

3.1 Participants and Domain

The study was conducted as part of the "Business Development with Information Systems – BDIS" course. BDIS is a 5 ECTS course, typically offered in the third semester of the "Bachelor's Degree Programme in Economics and Business Administration" in the Department of Management. The course is taught in English and it is designed to train students to analyze, evaluate, and apply models of information systems, decision making, and business management into the context of a comprehensive, semester-long case-study. The lecture material (i.e., slides, literature, etc.) is made available online one week in advance. Students are expected to read relevant literature and the lecture slides before coming to the class. To pass the course, students have to work in small groups, hand in a group case report, and pass an individual oral examination that includes questions related to the case and the conceptual knowledge of the domain.

Each year, approximately 180 sophomore students enroll in the course. Lectures are given weekly in an auditorium and last 2 hours. However, since lecture attendance is not mandatory, the number of students in the classroom varies each week. The study activity was available to all attending students. It is worth noting, though, that the study findings are based only on the sample of students that attended the course during all weeks of the study duration. Students attending the course only in some of the classes were also allowed to participate, but their data were not included in data analysis. Thus, only a total of 91 students participated in the study. Students were randomly distributed by the system into four groups, according to the feedback they were receiving during the revision phase (see next section). Student distribution into the four groups was:

- Control: 27 students;
- Confidence: 22 students;
- Preparation: 22 students;
- Both: 20 students.

Students volunteered to participate in the activity, which was not part of the official course assessment.

3.2 The SAGA System

The study used the "Self-Assessment/Group Awareness – SAGA" online quiz system. SAGA was developed by the research team of this study. Having a tailored-made system allowed for greater degree of flexibility in customizing the study variables and monitoring student activity. The system can provide the type of formative feedback that is not present in other quiz systems, while the ability to change its functionality allows the research team to use this system in a series of studies in different contexts and for different research purposes.

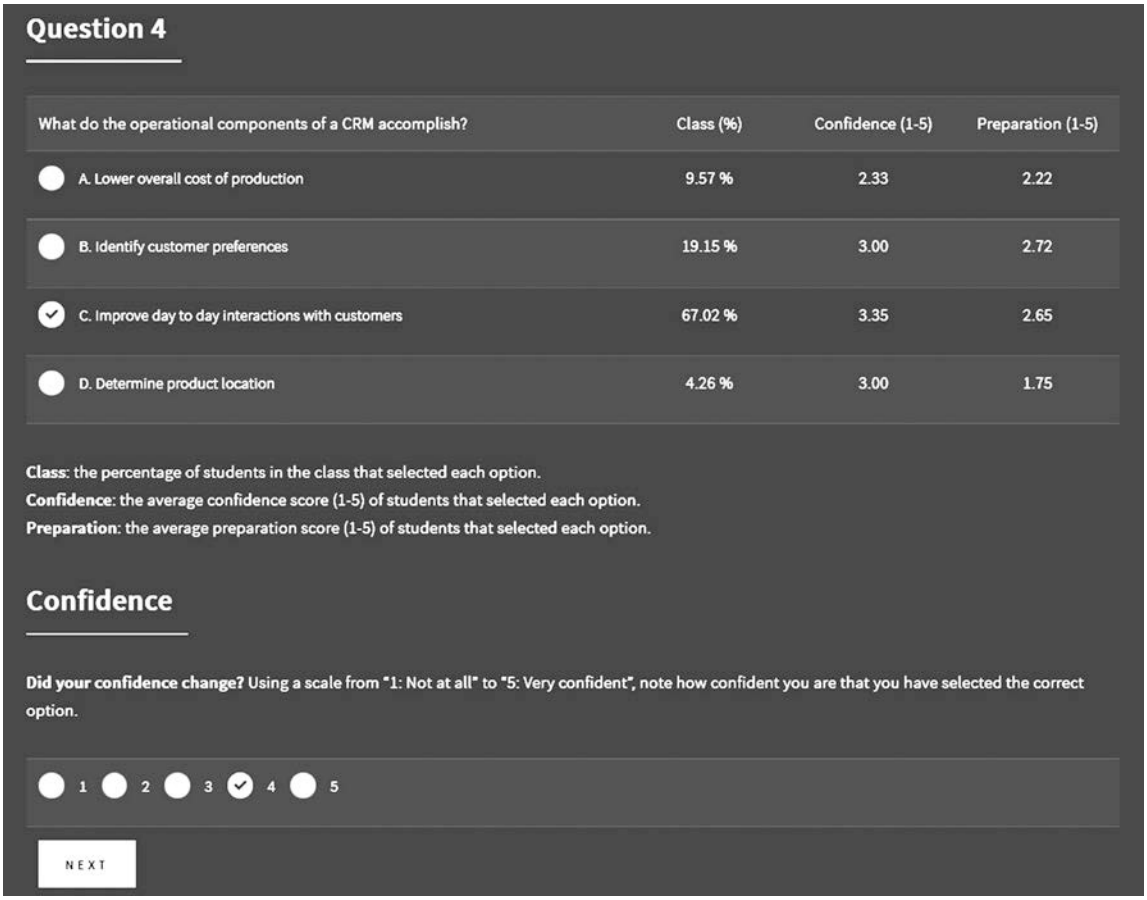

Figure 1: Screenshot of the SAGA system during the revision phase for students in the Both group (all metrics (percentage, confidence, and preparation are available).

Before the quiz activity, students have to answer a question in the system about the amount of time they spent preparing for the day's lesson: *"Some of the teaching material for today's class became available during the last week. Using a scale from '1: Not at all' to '5: I have read it thoroughly', how much time did you spent preparing for today's class?"*. Next, there is a series of eight multiple-choice questions prepared by the teacher, with four choices each. Each quiz question is accompanied by a question on students' confidence: *"Using a scale from '1: Not at all' to '5: Very confident', note how confident you are that you have selected the correct answer."*. Answering all questions (and their accompanying "confidence" questions) is mandatory. The questions are answered sequentially and the initial phase of the quiz ends, when the eighth question is answered.

In the revision phase that follows, students can browse through the eight questions and change, if they want to, their initial answers. For each of the four groups, SAGA provides a different set of information, based on the whole class population, to help students decide whether they should change their initial answers or not (Figure 1):

- Control: the percentage of student in the class that selected each option;
- Confidence: the percentage and the average confidence score of students that selected each option;
- Preparation: the percentage and the average preparation score of students that selected each option;
- Both: the percentage, the average confidence, and the average preparation scores of students that selected each option.

After the completion of the revision phase, the students are able to see their score and the correct answers. The teacher is able to monitor student progress and start the next phase of the quiz, when all students have finished the initial phase. It is

important to have all students on the same phase to ensure that all participants in the same study conditions will receive the same feedback from the system.

All students are in the same phase simultaneously during the activity and SAGA provides monitoring functionalities to the teacher, who is responsible for activating the next phase in the process.

3.3 Procedure and Study Conditions

The study was conducted during the Fall semester of 2016. The total duration of the study was five weeks, split into two parts, namely: the weekly quizzes (first four weeks) and the retention test and the final survey (fifth week).

During the first four weeks students in the BDIS course started the class by going through the three phases of the SAGA system (i.e., provide answers in the initial phase, change the answers during the revision phase, and see score and correct answers). The students were informed about the research nature of the activity and about the fact that they may receive different feedback information in the system than their fellow students.

A weekly quiz activity was designed to last up to 20 minutes, not to disrupt the lecture plan of the teacher. It needs to be emphasized, that the current study is part of a larger research project that explores the potential of educational technology in an efficient way for the teacher. This means that the planned activity should be able to enhance learning and engagement in a course, without increasing the workload overhead for the teacher and without taking too many resources (e.g., teaching or preparation time). According to the plan, students were given ten minutes to provide their initial answers, five minutes to revise them (optionally), and five minutes to discuss the correct answers with the teacher. After the quiz activity, the lecture proceeded as usual.

In the fifth week, students had to take an unannounced retention quiz and provide their input in a questionnaire recording their opinions and attitudes towards the whole activity. The retention test included four questions from the day's lesson and 16 questions that were previously included in the weekly tests during the first four weeks. Because of the length of the quiz, and since the goal was to measure retention, there was no revision phase and students skipped directly after the quiz to the correct answers and their scores. The questionnaire included open and closed-type questions, asking students to share their opinions about the helpfulness of the different feedback information they received, the impact of the weekly quizzes on their preparation strategies, and their suggestions for improvement.

The whole activity was individual and anonymous. No personal information about the students was recorded by SAGA, the researchers, or the teacher. The study conditions were identical for the four groups, except for the type of feedback they were receiving during the revision phase and the slightly different set of questions included in the final questionnaire.

3.4 Research Design

The study employed a between-subjects 2x2 factorial design with the study conditions in each group (i.e., type of feedback information, in addition to the percentage that was available to all students) being the independent variables (Table 1).

Table 1: Levels of independent variables and student groups.

		Confidence Feedback	
		No	***Yes***
Preparation Feedback	***No***	Control	Confidence
	Yes	Preparation	Both

Students' performance in the initial and the revision phase of the quiz (and the respective improvement recorded) throughout the five weeks and their responses in the questionnaire were the dependent variables of the study.

3.5 Data Collection and Analysis

A level of significance at .05 was chosen, for all statistical analyses. The study used parametric tests for the analysis of student performance and non-parametric tests for the analysis of student responses in the questionnaire, because for some of the examined variables the normal distribution criterion was violated.

Student performance analysis during the four quizzes was performed in two steps. First, a comparison between the groups in the four weekly quizzes was performed, taking into account all used questions (i.e., 32 in total; eight in each weekly quiz). In the second step student performance analysis focused only on a subset of the 32 questions. This analysis was conducted right after the fourth week of the study. The reason for such an approach was that it was not possible to identify during the design time of the study the challenging questions in which the feedback information that

Table 2: Student performance in the weekly quizzes, the subset of the 13 challenging questions, and the retention test. Scales – Weekly quizzes: 0-8; Challenging: 0-13; Retention: 0-16.

	Control			Confidence			Preparation			Both		
Week 1	M	*SD*	n	M	*SD*	n	M	*SD*	n	M	*SD*	n
Initial	4.58	*(1.53)*	27	4.48	*(1.19)*	22	4.15	*(2.34)*	22	4.85	*(1.73)*	20
Revision	6.25	*(1.32)*	27	6.40	*(1.29)*	22	5.62	*(2.22)*	22	6.46	*(1.39)*	20
Week 2												
Initial	3.50	*(1.27)*	27	3.64	*(1.17)*	22	4.13	*(1.14)*	22	4.06	*(1.43)*	20
Revision	4.35	*(0.87)*	27	4.01	*(1.34)*	22	4.69	*(0.94)*	22	4.50	*(1.04)*	20
Week 3												
Initial	5.52	*(1.64)*	27	5.19	*(1.74)*	22	5.43	*(1.59)*	22	5.09	*(1.63)*	20
Revision	6.87	*(1.10)*	27	7.08	*(1.38)*	22	7.00	*(1.00)*	22	7.05	*(1.25)*	20
Week 4												
Initial	3.73	*(1.98)*	27	3.52	*(1.37)*	22	4.14	*(1.83)*	22	4.05	*(1.43)*	20
Revision	5.76	*(1.04)*	27	5.26	*(1.05)*	22	6.14	*(1.15)*	22	6.06	*(1.21)*	20
Challenging*												
Initial	4.44	*(4.34)*	27	3.82	*(3.59)*	22	5.27	*(3.98)*	22	4.40	*(2.87)*	20
Revision	4.00	*(4.29)*	27	4.90	*(3.00)*	22	6.36	*(4.22)*	22	6.60	*(3.73)*	20
Retention												
Initial	10.00	*(3.23)*	27	10.86	*(2.14)*	22	10.68	*(3.24)*	22	10.80	*(3.2)*	20

* $p<0.05$

was given additionally to the percentage (i.e., confidence and preparation level) would be helpful. In other words, in case of an easy question, it was expected that a great student majority would have selected the correct choice during the initial phase of the quiz. As such, a high percentage value during the revision phase would have only provided reassurance to the students, suggesting that no revision is necessary.

Percentage is a commonly used metric in quiz systems and the expectation during the design of this study was that students in SAGA would be relying firstly on the percentage, before considering the information provided by the other two metrics. Following this line of argumentation, the claim was that the impact of the confidence and preparation feedback would only be observed in cases where the percentage alone could not "clearly" point at the correct choice.

The definition used in the study to identify these "clear" cases included three conditions that had to be true at the same time:

- The correct choice was also the most selected;
- The correct choice was selected by at least 50% of the students;
- The correct choice had a least 20 points difference from the second most selected choice.

In all other cases, the percentage information was considered either misleading (i.e., pointing at a wrong choice) or ambiguous (i.e., not pointing clearly at one choice). By applying this definition, the analysis revealed a subset of 13 challenging questions (four from the first, five from the second, one from the third, and three from the fourth week). Thus, the impact of confidence and preparation feedback on students' performance during the first four weeks was analyzed against this subset.

The retention test was designed after the fourth week and was compiled by (a) four new questions addressing the lesson on the fifth week (for this this reason, these four questions were not considered while measuring retention – they were included only because students were expecting questions for the day's lesson), (b) the 13 challenging questions, and (c) three additional old questions that were close to be categorized as challenging, in order to balance the number of questions from each weekly quiz (four from the first, five from the second, three from the third, and four from the fourth week).

4 RESULTS

Table 2 presents students' performance in the four weekly quizzes, the subset of the 13 questions that were identified as challenging, and the retention test that included the 13 challenging questions, plus three more, and was conducted on the fifth week.

4.1 Weekly Quizzes

As it is evident, students' performance each week varied, suggesting variations on their preparation level or the difficulty level of the topics covered through the questions. Two-way analysis of variance showed that the four groups performed similarly in the initial phase of the quiz in all first four weeks ($p > 0.05$). Two-way analysis of covariance, using students' scores in the initial phase of the quiz as a covariate, showed that the groups were also comparable in the final score (i.e., revision phase) in all four weekly quizzes ($p > 0.05$). In addition, paired-samples t-test results reveled that students in all groups improved their scores significantly from the initial to the revision phase, in all four weekly quizzes.

4.2 Subset Performance

Question analysis revealed that students relied strongly on the percentage metric. Applying the definition about "clearly" pointing to a correct choice, analysis showed that the percentage metric was pointing at a specific choice in 24 out of the 32 questions, during the initial phase of the quiz. However, only 19 of these choices were actually the correct ones, suggesting that many students that consulted the percentage metric revised their answers in these five questions to a wrong choice, trusting the majority of the class that did the same. The remaining 13 questions formed the challenging subset that was mentioned previously.

By applying a similar definition for "clearly" pointing at the correct choice for the confidence and preparation metrics, the analysis revealed that the confidence and the preparation metrics were pointing at the correct choice in 8 and 7, respectively, of the 13 challenging questions, in which the percentage metric was ambiguous or misleading.

Paired-samples t-test results showed that in the 13-question subset Confidence ($t[21] = 2.324$, $p = 0.030$, $d = 0.720$), Preparation ($t[24] = 2.027$, $p = 0.046$, $d = 0.630$), and Both ($t[19] = 2.979$, $p = 0.008$, $d = 0.970$) groups scores improved significantly during the revision phase, while the Control group was the only one that did not improve (getting slightly worse scores during the revision phase). Two-way analysis of covariance, using students' scores in the initial phase as a covariate, showed a significant main effect for the confidence ($F(1,86) = 4.115$, $p = 0.046$, $\eta^2 = 0.046$) and preparation ($F(1,86) = 7.153$, $p = 0.009$, $\eta^2 = 0.077$) metrics, but not for their interaction ($p > 0.05$).

4.3 Retention Test

Two-way analysis of variance showed that students in all four groups performed similarly in the 16 old questions that were included in the retention test ($p > 0.05$).

4.4 Student Opinions and Behavior

Table 3 presents students' responses in the most important items of the final questionnaire. Kruskal-Wallis and Mann-Whitney test results showed no significant differences in the responses of the four groups ($p > 0.05$). According to students' opinions, the most useful feedback metric for them was percentage metric ($M = 3.62$, $SD = 1.01$), followed by the confidence level ($M = 3.32$, $SD = 1.20$), and the preparation level ($M = 2.64$, $SD = 1.43$). In addition, students were asked to state their preference on additional types of feedback that are considered for future studies with SAGA: confidence ($M = 3.35$, $SD = 1.11$), past performance ($M = 3.20$, $SD = 1.14$), preparation ($M = 3.15$, $SD = 1.19$), argumentation ($M = 3.15$, $SD = 1.15$), and peer communication ($M = 2.87$, $SD = 1.19$), sorted from most to least desirable. The questions about confidence and preparation were addressed only to appropriate groups. Past performance referred to the average past scores (based on previous weeks) of students that selected each option; argumentation referred to a short argument for each option, written by an anonymous fellow student; and peer communication referred to the opportunity to briefly text anonymously with fellow students.

In Q1, students were asked whether the weekly quizzes increased the amount of preparation time each week. No significant difference was measured between the groups ($p > 0.05$), with students being split in their answers ($M = 2.49$, $SD = 1.28$). What is interesting though is that according to students' answers on the preparation question in the beginning of the quiz each week, it appears that students did increase the time they spent preparing for the course. Figure 2 presents the mean values for the preparation level each week for whole participant population. The results of the analysis of variance, with repeated measures with a Greenhouse-Geisser correction (sphericity assumption was violated), showed that the mean value for the preparation level

Table 3: Student responses in the questionnaire. Scale – 1: Not at all; 5: Very much.

	Control *n = 27*		Confidence *n = 22*		Preparation *n = 22*		Both *n = 20*		Total *n = 91*	
	M	*SD*	M	*SD*	M	*SD*	M	*SD*	M	*SD*
Q1. Has the quiz made you spent more time preparing during the week for each lecture?	2.17	*(1.04)*	2.90	*(1.17)*	2.68	*(1.39)*	2.17	*(1.37)*	2.49	*(1.28)*
Q2. Do you find the percentage values you see useful in choosing your final responses?	3.72	*(0.89)*	3.43	*(0.87)*	3.73	*(1.12)*	3.61	*(1.15)*	3.62	*(1.01)*
Q3. Do find the confidence values you see useful in choosing your final responses?	-	-	3.33	*(1.19)*	-	-	3.30	*(1.25)*	3.32	*(1.21)*
Q4. Do find the preparation values you see useful in choosing your final responses?	-	-	-	-	2.59	*(1.53)*	2.70	*(1.39)*	2.64	*(1.44)*
Q5. How useful do you think the confidence level (confidence level of fellow students that selected each option) would be for you in choosing your final answers?	3.61	*(0.85)*	-	-	3.14	*(1.28)*	-	-	3.35	*(1.12)*
Q6. How useful do you think the preparation level (average preparation level of fellow students that selected each option) would be for you in choosing your final answers?	3.28	*(1.22)*	3.05	*(1.20)*	-	-	-	-	3.15	*(1.20)*
Q7. How useful do you think the past performance (average past scores – based on previous weeks – of fellow students that selected each option) would be for you in choosing your final answers?	3.83	*(0.85)*	2.95	*(1.28)*	3.14	*(1.28)*	2.00	*(0.95)*	3.20	*(1.14)*
Q8. How useful do you think argumentation (a short argument for each option, written by a fellow student – anonymity remains) would be for you in choosing your final answers?	2.72	*(1.36)*	3.05	*(0.97)*	3.18	*(1.25)*	3.22	*(1.04)*	3.06	*(1.15)*
Q9. How useful do you think peer communication (opportunity to briefly text anonymously with fellow students) would be for you in choosing your final answers?	2.78	*(1.06)*	2.95	*(1.16)*	2.95	*(1.49)*	2.78	*(1.08)*	2.87	*(1.20)*

were statistically significantly different ($F(3.306, 247.966) = 44.128$, $p = 0.00$, $\eta^2 = 0.370$).

Pearson correlation coefficient test results showed that confidence, preparation, and initial performance scores were all significantly correlated ($p < 0.01$) throughout the four weeks, suggesting that students that felt confident and prepared were, indeed, performing better in the weekly quizzes. In addition, paired-samples t-test results showed that students' confidence increased significantly ($p < 0.05$) from the initial to the revision phase of the quiz, for all groups, in all four weeks, in which revision phase was available.

In the open-ended items of the questionnaire, students commented positively on the activity ("*Nice program design, well-put questions.*"; "*The quiz is a good starting point for the lectures, however they should be kept short.*"; "*I really like that you asked us about these things. I am a huge fan of giving feedback and striving for improvement. I am a highly competitive person and the quizzes are compelling to me.*").

Regarding suggestions for improvement in future implementation of SAGA, students suggested gamification ("*Maybe a leaderboard/high score list.*"), information on the wrong answers ("*It might be nice to know which answers we already got wrong.*"), additional information on peers ("*How*

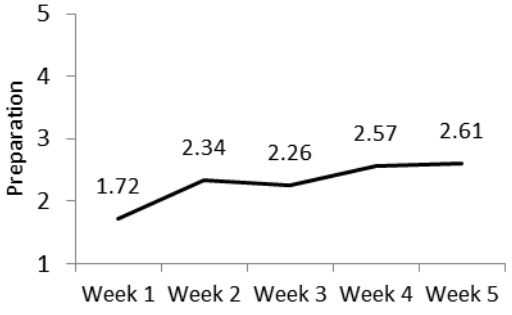

Figure 2: Student preparation values.

many lectures the persons have participated in."), feedback from a specific group of people ("*my study-groups feedback.*"), and splitting the two phases of the quiz before and after the lecture ("*Reading the actual curriculum before the class OR repeat the second phase* [i.e., revision] *of the quiz at the end of the class to actually see if we are taking something out of the lecture.*").

Finally, it is worth mentioning that although the a weekly quiz activity was designed to last up to 20 minutes, SAGA log files revealed that students needed on average six minutes for the initial phase and 4 minutes for the revision one. This gave more time to the teacher, who was able to spend more time on discussing the questions and revisit them during the lecture, when related material was presented.

5 DISCUSSION

Results analysis showed that when taking into account a whole quiz, student performance is comparable in the four groups. As explained, this could have been expected, since the need for additional feedback increases in cases of uncertainty and ambiguity. The number of such cases in the weekly quizzes could not be foreseen. A series of factors such as student preparedness level, difficulty and complexity of course topics, expectancy of certain questions could all affect students' performance in the initial phase of a quiz, leaving either too much or too little space for considering revisions. The small number of challenging questions that would require additional support to the students even out any observed differences, while can only hypothesize that significant differences may be discovered in longer quizzes.

The absence of significant differences in the retention test can be easily explained by the fact that the weekly quizzes were administered in the beginning of the lesson and the teacher had the remaining of the two hours to present the day's topics and resolve any misconceptions revealed by examining students' quiz performance. In this way, it can be argued that the quiz served its instructional purpose by making misconceptions obvious and allowing the teacher to tailor the lecture accordingly. It is worth noting that the average score the four groups achieved in the retention test is considered satisfactory (with 10% of the student population achieving a perfect score), especially since some of the questions were related to topics that had been covered a month ago.

Case-by-case analysis showed that students relied heavily on the percentage metric in identifying the correct answer and it was revealed, they did so, even in cases where the suggested choice was wrong. Despite that, the percentage metric still remains a commonly used way to provide a picture of a group's position on an issue and this study is not arguing, of course, for the abandonment of this metric. The percentage metric is objective, easily understood, and satisfactory in indicating the correct answer (19 out of 32, in this study). However, what it is argued in this study is that the percentage does not carry any information about the people that are behind the figures, and this information may be vital, in cases where the population is split.

Confidence and preparation metrics, on the other hand, provide qualitative information on the participants, but they both rely on participants' metacognitive level and their ability to accurately assess their preparation and confidence levels. In the current study, both metrics were significantly correlated to the initial performance, suggesting that they could both indicate adequately the correct answer. The question that rises, though, is whether students appreciate these metrics are useful and if they base their activity on them. In the questionnaire, students evaluated positively the percentage and confidence metrics, while they were split about the preparation one. One reason for this may be that students value the confidence metric more because it provides a picture of peers' understanding *after* a question was answered, while the level of preparation is noted in the beginning of the activity, *before* any of the quiz questions becomes available.

Nevertheless, student performance analysis on the subset of questions in which percentage could not provide enough support clearly revealed that all treatment groups outperformed the Control group. This finding provides evidence on how simple metrics, such as the confidence and preparation, could be easily integrated in quiz activities and enhance student performance.

Regarding student behaviors and attitudes towards the activity, students' increased level of preparation throughout the study duration is a very positive indication of the kind of impact such quiz activities could have on student engagement in the course. This increase on preparation time is not attributed to a specific study condition and it apparent in all groups.

According to students' statements, the activity was positively received and several of the suggestions for improvement are already included in

design of planned studies. Regarding feedback types that could be added in SAGA, peer confidence was the most desirable option amongst the students in the Control and Preparation groups. Students' past record came second, suggesting that students are in favor of objective metrics, even though, good past performance does not guarantee high performance in a new topic. It is worth noting that, although they were the least desirable, reading an argument for each group choice and directly texting anonymously with a peer were both evaluated positively.

6 CONCLUSIONS

The study provided useful evidence on how additional subjective metrics could complement an objective metric, such as the percentage, and provide better support to students in multiple-choice quiz activities. The implications for designers and teachers that use quiz tools suggest that metrics that would better describe the participants are easy to use and have a significant effect on students' performance. The level of confidence and preparation (in addition to the other scaffolding methods mentioned in the questionnaire) could be translated to questions an individual could ask himself/herself about his/her peers: What do the others say (percentage)? How good are they (past performance)? How much have they studied (preparation)? Why did they say that (argumentation)?

Future studies will focus on additional metrics, addressing also some of the limitation in this study. As such, future studies are planned with larger audiences, different subject matters, and multimodality in representation of the metric information (e.g., combination of text with graphs and color schemes). Finally, as it was already mentioned, another side of this series of studies is focusing on the effect these shorts quizzes could have on student engagement and performance in the course. A future study is planning to compare classes with and without the quiz activities.

ACKNOWLEDGEMENTS

This work has been partially funded by a Starting Grant from AUFF (Aarhus Universitets Forskningsfond), titled "Innovative and Emerging Technologies in Education".

REFERENCES

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why Students Engage in "Gaming the System" Behavior in Interactive Learning Environments. *Journal of Interactive Learning Research. 19(2)*, 185-224.

Bodemer, D. (2011). Tacit guidance for collaborative multimedia learning. *Computers in Human Behavior, 27(3)*, 1079–1086.

Bransford, J. D., Brown, A., & Cocking, R. (2000). *How people learn: Mind, brain, experience and school*. Washington, DC, National Academy Press.

Buder, J. (2011). Group awareness tools for learning: Current and future directions. *Computers in Human Behavior, 27*, 1114–1117.

Buil, I., Catalán, S., & Martínez, E. (2016). Do clickers enhance learning? A control-value theory approach. *Computers & Education, 103*, 170-182.

Denny, P. (2013). The effect of virtual achievements on student engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13*). ACM, New York, NY, USA, 763-772.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining "gamification". *In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*. ACM, New York, 9-15.

DiBattista, D., Mitterer, J. O., & Gosse, L. (2004). Acceptance by undergraduates of the immediate feedback assessment technique for multiple-choice testing. *Teaching in Higher Education, 9(1)*, 17-28.

Erkens, M., Schlottbom, P., & Bodemer, D. (2016). Qualitative and Quantitative Information in Cognitive Group Awareness Tools: Impact on Collaborative Learning. In Looi, C.-K., Polman, J., Cress, U., & Reimann, P. (Eds.), *Transforming Learning, Empowering Learners: 12th International Conference of the Learning Sciences* (pp. 458-465). Singapore: International Society of the Learning Sciences.

Janssen, J., & Bodemer, D. (2013). Coordinated computer-supported collaborative learning: Awareness and awareness tools. *Educational Psychologist,48*, 40–55.

Kleitman, S., & Costa, D. S. J. (2014). The role of a novel formative assessment tool (Stats-mIQ) and individual differences in real-life academic performance. *Learning and Individual Differences, 29*, 150-161.

Lin, J. -W., Mai, L. -J., & Lai, Y.-C. (2015). Peer interaction and social network analysis of online communities with the support of awareness of different contexts. *International Journal of Computer-Supported Collaborative Learning, 10(2)*, 139-159.

Méndez-Coca, D., & Slisko, J. (2013). Software Socrative and smartphones as tools for implementation of basic processes of active physics learning in classroom: An initial feasibility study with prospective teachers. *European Journal of Physics Education, 4(2)*, 17-24.

Papadopoulos, P. M., Demetriadis, S. N., & Weinberger, A. (2013). "Make It Explicit!": Improving

Collaboration through Increase of Script Coercion. *Journal of Computer Assisted Learning, 29 (4),* 383 – 398.

Papadopoulos, P. M., Lagkas, T. D., & Demetriadis, S. N. (2016). How Revealing Rankings Affects Student Attitude and Performance in a Peer Review Learning Environment. *Communications in Computer and Information Science (CCIS): Computer Supported Education 2015. Vol. 583* Springer Verlag, 2016. p. 225-240.

Schnaubert, L., & Bodemer, D. (2015). Subjective Validity Ratings to Support Shared Knowledge Construction in CSCL. In O. Lindwall, P. Häkkinen, T. Koschmann, P. Tchounikine, & S. Ludvigsen (Eds.), *Exploring the Material Conditions of Learning: The Computer Supported Collaborative Learning (CSCL) Conference 2015 (Vol. 2)* (pp. 933-934). Gothenburg: International Society of the Learning Sciences.

Sosa, G.W., Berger, D. E., Saw, A. T., &Mary, J. C. (2011). Effectiveness of computer-assisted instruction in statistics: A meta-analysis. *Review of Educational Research, 81(1),* 97–128.

Wang, A.I. (2015). The wear out effect of a game-based student response system. *Computers & Education, 82,* 217-227.

Wang, T.-H. (2008). Web-based quiz-game-like formative assessment: Development and evaluation. *Computers & Education, 51,* 1247-1263.